Bannir les IA voyous

Dominique Boullier

Le 19 Décembre pour AOC

Le discours généraliste sur l'IA, voire sur les IA ne porte guère sur des systèmes d'IA précis (avec toutes leurs spécificités logicielles : ce n'est pas la même chose d'utiliser Copilot et ChatGPT grand public même si la techno est voisine) ni sur les firmes de l'IA. Pour ces dernières, on s'enflamme avant tout sur leurs levées de fonds, sur leurs performances mais on globalise vite pour parler de « la Tech » sans entrer dans les détails, ce qui permet toutes les variations à prétention philosophique voire prophétique. Quand on prétend s'y intéresser de plus près, ce sont avant tout leurs discours, leurs inspirateurs, leurs positions politiques qui fascinent comme source d'une nouvelle idéologie dominante, libertarienne/ fasciste pourrait-on presque dire. Et du côté des utilisateurs, des professionnels de l'IA, on subit l'avalanche de benchmarks, de tests de capacités nouvelles dans des situations souvent mal identifiées ou peu soucieuses de repérer les fonctionnements réels des organisations. Tout cela encourage les discours généraux qui font oublier qu'il existe d'autres IA que l'IA générative, qu'il existe même des usages alternatifs des IA, qu'il existe des chercheurs qui continuent à exiger de l'explicabilité par exemple, qui développe des systèmes d'IA plus petits, etc. Lorsque j'alerte sur les « déshumanités numériques » (Armand Colin, 2025), je ne veux en aucun cas participer au chœur des lamentations des prophètes qui grossissent tellement l'affaire qu'on se retrouve chacun impuissant. Oui, on peut faire autrement ! Oui, on peut aussi attaquer ces firmes, d'autant plus lorsqu'elles en sont à leurs débuts, plus fragiles et plus arrogantes, c'est ce que je propose ici.

C'est pourquoi j'ai tenté d'inverser cette tendance aux généralités sur l'IA-qui-fait-peur pour parler de « systèmes d'IA » et non d'IA et insister sur le pluralisme de ces systèmes dont notamment l'IA symbolique qui a permis de maintenir la prise sur le monde (voir article AOC). Mais il faut sans doute aller plus loin et frapper plus fort. Même en acceptant d'utiliser des LLM (Large Language Models), dont l'avenir est très sombre, financièrement comme techniquement, il faut désormais écoper en urgence en éliminant du secteur des firmes malfaisantes, des « IA voyous » pour faire simple.

Oui, il faut bannir les firmes qui ont mis sur le marché (tout cela sans autorisation évidemment) des produits nocifs, addictifs, à risque, et qui s'étendent désormais à tous les domaines, comme on le ferait dans n'importe quel secteur industriel. Or, toutes les plateformes ne pratiquent pas les mêmes méthodes (car toutes les grandes firmes d'IA ont ou prétendent avoir à terme un statut de plateforme, point de passage obligé pour tous les versants des transactions). Cela ne blanchit en rien les autres ni ne les dédouane de leurs responsabilités, d'autant plus grandes qu'elles sont très puissantes, adossées à des plateformes d'influence inédites. Mais en raison de leurs impératifs de

réputation, elles ne peuvent se permettre de jouer la provocation trop souvent. Elles doivent donc rester prudentes tout en s'attachant à suivre le rythme de disruption introduit par les nouveaux entrants de façon à continuer à attirer des investisseurs, puisque tout est ici un concours de cash à brûler pour des fermes de serveurs toujours plus énormes.

Name and shame : OpenAI et xAI

Deux d'entre elles doivent être nommées et blâmées (name and shame) pour entrainer ensuite leur bannissement de tout le secteur et de toute interaction avec le grand public et avec les marchés publics en particulier. J'ai nommé OpenAI (ChatGPT) et xAI (Grok). Voilà les deux IA voyous qui donnent le rythme de la spéculation financière et des outrances des contenus si fréquentes qu'elles pourraient se banaliser et paralyser les gouvernements. Elles semblent dans les deux cas avoir franchi des limites ces dernières semaines, même si la notion de limite est précisément une notion qu'elles ignorent, toutes emportées par l'hubris qui caractérise leur impératif supérieur de disruption.

Cela rappelle étrangement les méthodes et les principes d'Uber. Musk et Altman tiennent la comparaison avec Kalanick, le CEO délirant et délinquant qui a dû être débranché par les VC (Venture Capitalists) de la Silicon Valley tant il semait la guerre partout comme le veut le dogme disruptif qui est le leur. Les personnalités de ces disrupteurs ne m'intéressent guère et devraient cesser de faire la une des journaux voire des travaux des chercheurs car on perd de vue alors l'entreprise systémique de destruction de l'état de droit qui est au cœur des pratiques des firmes dans leurs choix techniques mêmes. Car c'est la menace essentielle : une aliénation générale des consommateurs à la gloire de délinquants sans foi ni loi qui adoptent, au choix, les méthodes des voyous, des dealers ou des mafieux, selon les circonstances.

Du point de vue réglementaire, je propose de les nommer des « IA voyous » pour faire référence aux « Etats voyous » (Rogue States), terme qui s'est imposé à l'international sous l'influence américaine et qui a entrainé à diverses époques et pour diverses raisons le bannissement de certains états de toute relation internationale, puisque leurs activités étaient délibérément orientées vers le sabotage de ce droit lui-même. Certains diront sans doute que ce qualificatif aurait pu et pourrait alors s'appliquer très largement à des Etats qui n'ont que faire de leurs engagements internationaux et du droit international et les exemples abondent en ce moment en effet. Le terme date sans doute d'une époque pré 2022 où l'on pouvait encore feindre de croire à l'existence de ces règles juridiques dépassant les Etats. 2022, année fatale, qui voit à la fois l'invasion de l'Ukraine par la Russie, le rachat de Twitter par Elon Musk et le lancement de ChatGPT par OpenAI fin Novembre. 2022, peut sans doute constituer un jalon facile à mémoriser pour signaler un moment de bascule vers un retour du Far-West, pas

romantique du tout, mais avant tout prédateur (quand bien même la prédation existait bien avant mais s'efforçait encore de s'affubler des atours du droit).

Et les effets sur le débat public et les consciences collectives du monde entier ont été dévastateurs. On tend à mettre à part ce qui se joue dans l'espace médiatique en considérant que tout cela ne relève que d'images, de discours, et d'entertainment quasiment, à la différence d'une guerre comme en Ukraine. Pourtant, la destruction de l'esprit public à l'œuvre désormais ne peut être sous-estimée. Je l'ai analysée et combattue depuis l'introduction même de la publicité dans les réseaux sociaux à partir de 2009 et je l'ai encore combattue contre Uber dès 2013 et ses pratiques de disruption de bandit jouisseur du chaos qu'il entraine (voir les livres de Laurent Lasne à partir de 2015). Mais lorsqu'on croit toucher le fond avec Cambridge Analytica, on découvre que certains travaillent encore à aggraver notre détresse mentale collective, délibérément, pour attirer les investisseurs. C'est pourquoi depuis 2022, on peut parler d'IA voyous comme d'une cible nouvelle, entretenue par tout l'écosystème numérique qui est toujours, rappelons-le, un capitalisme financier numérique avec les principes et les mœurs des firmes spéculatives qui guident le monde désormais.

La liste des pratiques d'IA voyou (et des motifs de poursuite)

Prenons les pratiques de Open AI/ ChatGPT typiques d'un comportement d'IA voyou.

Un modèle d'affaire voyou en trompe l'oeil

- S'appeler OpenAI sans rien à voir avec l'open source, veillant au contraire à fermer et contrôler son code (ce que plusieurs autres firmes d'IA pratiquent aussi)
- S'arranger pour combiner un statut de fondation non-profit avec une activité commerciale à but très lucratif, dans la confusion la plus totale, entrainant le départ de plusieurs cadres (et aussi un litige avec Elon Musk, le champion bien connu de la probité dans les affaires).
- Revendiquer son indépendance et en fait bénéficier des investissements massifs de Microsoft dès son démarrage, au point de rendre opaques la nature des connexions avec les services d'IA de Microsoft comme Copilot exploitant largement les ressources de ChatGPT. Microsoft possède toujours 27% du capital de la Public Benefit Corporation créée par Open AI, dans un accord de partenariat jusqu'en 2025. Microsoft Azure offre ses capacités de calcul à Open AI.
- Organiser l'attrait des investisseurs autour d'une promesse de monopole, comme l'ont fait bien d'autres plateformes devenues des monopoles avant lui et respectant le dogme du libertarianisme, au mépris de la réalité des rapports de force sur le marché.

- Créer un système d'investissement circulaire avec des partenaires/ fournisseurs/clients sous forme de participations croisées (notamment avec Nvidia) rendant illisible toute la solvabilité réelle de l'entreprise, au point de relever d'une pyramide de Ponzi à ciel ouvert. Malheur à celui qui lâchera prise le dernier. Plusieurs avertissements récents (cf. Michael Burry) indiquent bien la crainte de l'éclatement de la bulle ainsi créée.
- Jouer comme d'autres avant lui sur l'absence de retour sur investissement et sur la seule valorisation du titre. L'absence de rentabilité (de ROI) actuelle devient même un argument pour montrer tout l'investissement fait et à faire encore. Open AI a perdu 12 Milliards de dollars au 3e trimestre 2025. OpenAI s'est pourtant engagée à 1,4 trillion de dollars de dépenses sur les prochaines années chez Oracle, NVIDIA, Microsoft, AMD, Broadcom, Amazon, ce qu'il faut comparer aux 20 milliards de revenus obtenus en 2025.
- Prétendre que ce montage opaque est devenu tellement vital pour toute l'économie étatsunienne que l'Etat fédéral devra aider la firme en cas de problème (Sarah Friar, la directrice financière) selon l'adage « too big to fail », puis retirer ces expressions pourtant publiées face au tollé suscité par ce chantage à l'Etat de la part d'une firme archi libertarienne.
- Appeler des fonds pour financer une croissance illimitée en matière de data centers, ce qui entrainerait des consommations électriques impossibles à satisfaire avec les ressources actuelles, privant ainsi des communautés locales des ressources aquifères et sans aucune prise en compte du réchauffement climatique. Le même enjeu étant amplifié par la consommation d'eau et la pollution systématique ainsi entrainée.
- Se présenter comme le leader du rush pour l'IA au point d'obtenir le soutien de l'administration Trump pour la création de super infrastructures supposées indispensables (projet Stargate avec Oracle et Softbank). Au même moment (janvier 2025) le chinois Deep Seek prétend pourtant obtenir les mêmes résultats avec dix fois moins de ressources de calcul. Sans valider trop vite cette dernière information sans examen précis, elle indique cependant que d'autres façons de faire, moins consommatrices de ressources, seraient possibles et déjà disponibles mais toujours maintenues dans l'ombre en cultivant la croyance et l'ignorance des utilisateurs et des investisseurs et des gouvernements.
- Organiser la riposte à Google dès lors que ses avancées avec Gemini 3 menace la réputation et l'attractivité de ChatGPT, au point de provoquer la comparaison avec Netcape, le premier navigateur du web disqualifié par Internet Explorer.

Un grand prédateur de données et de travail caché…qui signe des accords après coup

- Organiser la prédation systématique de toutes les données disponibles sur le web sans en avertir les ayants droits, provoquer une grève des artistes de

Hollywood, pour entrainer son IA ChatGPT à moindre coût. Non-respect de la propriété intellectuelle à tous les niveaux.

- Négocier, après les avoir pillés, des accords avec les titres de la presse qui leur offrent des services vagues, opaques et de qualité non garantie pour pouvoir continuer à exploiter leurs contenus. Le Monde a signé ainsi un accord, le New York Times a refusé pour finalement signer le même type d'accord avec Amazon. Pour mémoire et pour mesurer l'arnaque, voici la définition que donnait Le Monde de ChatGPT : « Pour le grand public, les effets de cet accord seront visibles dans ChatGPT, que l'on peut décrire, en termes simples, comme un moteur de réponses (output) composées à partir des faits établis ou des commentaires exprimés par un nombre limité de références. Ce moteur fournit la réponse synthétique la plus plausible et prédictive à une question donnée. » On mesure l'incompréhension, la naïveté et la distance avec ce que l'on sait maintenant des performances et des combines d'OpenAI.
- Annoncer des chiffres sur le nombre de clients invérifiables, confondant les accès gratuits et payants, les accès occasionnels ou de découverte et les comptes actifs.
- Utiliser massivement des travailleurs du clic pour entrainer ses modèles, contrairement aux annonces faites puisqu'il faut toujours superviser et valider les constituants de base des modèles fondationnels ou d'applications nouvelles plus spécifiques. Les conditions de travail et de rémunération de ces travailleurs précaires ont été dénoncées comme un esclavage moderne qui engagent la responsabilité sociale d'Open AI qui n'a jusqu'ici proposé aucune démarche dans ce sens.

Multiplier les services pour se vendre comme plateforme en toute irresponsabilité

- Avoir déstabilisé tout le secteur de l'éducation en permettant d'accéder à un assistant qui amplifie la triche au point de la considérer comme l'état normal de la compétition scolaire à laquelle il faudrait s'adapter. Ce qui revient à se former à la conduite automobile en étant averti que le système et ses utilisateurs se comportent comme des chauffards, sans respecter aucune règle (la version avertie de l'éducation à l'IA est ainsi une éducation au monde de brutes qui vient grâce à ChatGPT et donc une défaite éducative).
- Annoncer sans cesse de nouvelles versions de ChatGPT dont les améliorations sont douteuses voire inexistantes, notamment avec ChatGPT 5o. Premier véritable cas de versioning qui détériore l'expérience utilisateur, ce qui a entrainé des abandons importants de clients. Le degré zéro de la culture industrielle et de l'orientation client. Les annonces de versions ne servent qu'à rassurer les investisseurs.

- Multiplier les services périphériques et les annonces pour attirer de nouveaux publics et surtout des investisseurs : Sora pour générer des vidéos (et donc du deep fake), compagnon érotique, etc. au détriment de l'amélioration réelle de la qualité, de la fiabilité ou de l'efficience des modèles.
- Continuer à attirer des investisseurs avec ces innovations périphériques sans proposer de retour sur investissement avant des années.
- Lancer Sora qui met à disposition du grand public un générateur de deep fake, sans aucun garde-fou, sans contrôle des usages et des contenus, sans limite d'âge, en inondant le marché de productions illégales en matière de respect des droits de la personne, en intoxiquant l'espace public avec ces versions frauduleuses de situations fabriquées, au mépris de la vérité. Sans doute la plus grande faute de OpenAI qui doit être enregistrée comme possible « crime contre l'humanité », dont on mesurera les conséquences dans les toutes prochaines années, sans attendre une supposée AGI.

Un modèle de réponses qui se transforme en compagnon toxique

- Produire des réponses systématiquement (et non des aveux d'ignorance quand c'est nécessaire), ne pas pondérer ces réponses d'une marge d'erreur (calculable puisque ce sont des probabilités à toutes les étapes du process), ne pas fournir de sources identifiables clairement (seulement des sites de référence depuis quelques mois). Présenter ainsi des informations avec l'assurance de la fiabilité sans en fournir un seul élément de vérification pour l'utilisateur. Tromperie claire sur la marchandise.
- Offrir un Compagnon toujours empathique, jamais déplaisant ou contradicteur pour renforcer encore la bulle informationnelle qui rend les débats impossibles et affaiblit la curiosité et le désir d'exploration et de surprise, l'intérêt d'origine d'un moteur de recherche, désormais remplacé par un moteur de réponse non fiable mais complaisant (mais Google avait commencé à privilégier les réponses depuis 2013). Le capitalisme de surveillance se présente ainsi comme un capitalisme de la bienveillance.
- Permettre la connexion permanente avec son compagnon sur tous supports et dans toutes situations, de travail, domestique ou publique, en mélangeant les sphères et en encourageant l'immersion pour rendre incontournable ce nouveau point de passage obligé dans toutes les relations, une plateformisation individuelle, augmentée par les agents.
- Offrir désormais une version érotique du même compagnon, pour adultes certes (avec quel contrôle d'âge ?), mais qui va permettre de rendre stimulante et addictive cette relation qui vise la fusion. Le générateur de vidéo Sora permettra d'y ajouter tous les fakes possibles pour personnaliser et mettre en situation de compagnon Sora. Ignorer ainsi les alertes sur les risques psychiques de ces applications, pour les personnes souffrant de solitude notamment.

- Proposer d'introduire de la publicité dans cette relation avec une interface pour tenter de monétiser un service très consommateur de ressources et sans modèle économique pour l'instant. La confusion entre tous les contenus et l'absence de sourcing et de labellisation des contenus, habituelle chez OpenAI, entrainera une publicité masquée dans toutes les situations, très rémunératrice et très manipulatrice.
- Tenir un discours inquiétant sur l'IA générale avec pétitions et prises de parole choc pour éviter de parler des problèmes posés par l'IA générative actuelle, celle développée à marche forcée par OpenAI. Engendrer ainsi la confusion la plus totale sur le futur lointain et sur des responsabilités très lointaines, tout en continuant à alimenter la frénésie d'utilisation et d'investissement actuelle.
- Prétendre embaucher fin décembre 2025 un « chef de la préparation »pour encadrer les modèles au vu des problèmes observés dans l'année notamment l'impact sur la santé mentale et les problèmes de sécurité, alors que les responsables des mêmes missions ont abandonné en 2024 et 2025 car toutes les orientations du service vont dans le sens contraire, vers une accélération d'offres toujours plus risquées, non testées.

La responsabilité des états et la fin de la fascination masochiste pour la disruption

La liste pourrait encore être allongée sans doute. Mais il doit être clair que tout partenariat avec ce type d'IA, individuel ou institutionnel, est un pacte avec le diable, avec une entité sans foi ni loi, prête à mentir sans cesse pour attirer les investisseurs puisqu'il s'agit avant tout de promesses et non de retours sur investissements réels qui sont actuellement inexistants, malgré les avantages immédiats perçus dans les usages grand public. Les nuages annonciateurs d'éclatement de la bulle IA s'amoncellent et OpenAI sait très bien qu'il sera la première compagnie dans la tourmente, la dernière entrante et celle qui a fait grossir la grenouille plus que les bœufs voisins et qui risque donc d'éclater rapidement. Mais puisque la firme s'est drapée dans les oripeaux de la firme disruptive, qui prend de risques, elle, et qui casse l'ordre établi, elle est parvenue à séduire tous les masochistes des gouvernements, dépassés par cette supposée fatalité au nom de la tyrannie du retard.

Les gouvernements ne semblent ainsi pas s'inquiéter encore de ces comportements de voyous, de même que certains semblent découvrir qu'il est impossible de faire confiance à Facebook, malgré tout ce que l'on sait sur ces plateformes de réseaux sociaux depuis Cambridge Analytica au moins. OpenAI bénéficie toujours d'une aura de pionnier, même si ses paris commencent à inquiéter. Les gouvernements doivent prévenir la chute annoncée en visant et liquidant le maillon faible, la firme voyou qui a lancé toute cette course, comme on a liquidé Leman Brothers (mais en fait on a continué comme avant, il ne faut pas de leurrer).

L'autre IA voyou : Grok

Dans le même wagon, il faudrait aussi bannir Grok, l'IA de xAI, l'IA de Musk omniprésente et entrainée sur X, annoncée délibérément comme une arme de destruction massive du wokisme, qui s'est vite transformée en agent fasciste dans de multiples réponses négationnistes et donc illégales en France. Il faudrait faire ici la liste des délits, des fraudes, des mensonges, des calomnies que Grok diffuse en permanence et qui sont amplifiées par son couplage étroit avec X. Mais l'énergie me manque, je dois le dire. Les tares structurelles d'Open AI sont tout aussi présentes mais amplifiées par le discours explicitement réactionnaire du propriétaire de X. Les derniers épisodes ont montré par exemple que le « mode épicé » de Grok permet de générer des conversations érotiques qui en fait autorisent aussi des prompts tels que « mets-la en bikini » ou « enlève-lui ses vêtements ». Les réactions ont été vives du côté du gouvernement mais on espère que maintenant que le Président semble avoir compris que « les plateformes se moquent de nous », des mesures immédiates de contrôle et d'interdiction seront prises.

Etablir une responsabilité d'éditeur de contenus

Les systèmes d'IA génératives des firmes révèlent dans ce cas leur pouvoir éditorial, un pouvoir de médias dès lors qu'elles filtrent les visions du monde acceptables par le propriétaire de la plate-forme ou non. Leur statut ne peut rester celui d'une plateforme de service banale comme un service de météo. Le cahier des charges qui doit leur être appliqué devra se rapprocher de celui des médias, comme je l'ai proposé depuis 2020 pour les plateformes de réseaux sociaux (qui sont désormais étroitement imbriquées avec l'IA). Cela demande une élaboration du cadre légal plus fine que celle que nous connaissons actuellement, qui se résume à l'AI Act européen, et qui empêcherait cette possible dérive. Rien n'interdit aux états d'augmenter ce règlement de versions plus contraignantes et de profiter notamment de toutes les occasions de litige causés par les comportements explicitement voyous de ces firmes pour les trainer en justice.

Une offensive légale et collective est nécessaire

Pour parvenir à cela, il faut mettre en place des opérations de tests de ces IA voyous, en posture d'utilisateur ordinaire pour démontrer publiquement et rigoureusement leurs dérives dans leurs réponses et les risques que cela entraine. Ces risques font partie du processus des « déshumanités numériques » que j'ai analysé dans mon livre récent. C'est dire que les enjeux de cette désintoxication précoce sont essentiels pour les architectures de connaissance et de vie publique que nous allons exploiter pendant les années à venir. Pour cela, on peut faire appel à des équipes de chercheurs qui sont déjà nombreux à organiser des formes de benchmarks/ défis pour voir comment faire déraper ces IA. Mais plutôt que de le faire à de fins de publication ou d'entertainment ou d'expérience esthético-philosophique, il serait temps d'organiser une procédure complète en référence à un cadre légal à multiples facettes car tous les comportements

de ces plateformes d'IA remettent en cause l'état de droit. Mais les contributions des internautes à ce repérage, à ces tests et à leur dénonciation pourrait même devenir un challenge, qui pourrait par exemple être coordonné par le Chaos Computer Club qui a appelé à quitter les applications des GAFAM lors de sa dernière conférence.

Dans ces cas précis des IA voyous, il sera bien difficile d'argumenter l'indulgence au nom de l'impératif d'innovation (qui est ici clairement socialement inutile et même nuisible), ou au nom de la liberté d'expression qui n'est que le masque de la désinhibition générale porteuse d'un monde invivable, de la loi des pulsions du plus fort. Comme je l'avais déjà observé en 1986 avec la Citizen Band : « Connecter n'est pas instituer » et l'on sait désormais que ces firmes sont avant tout des machines à désinstituer.